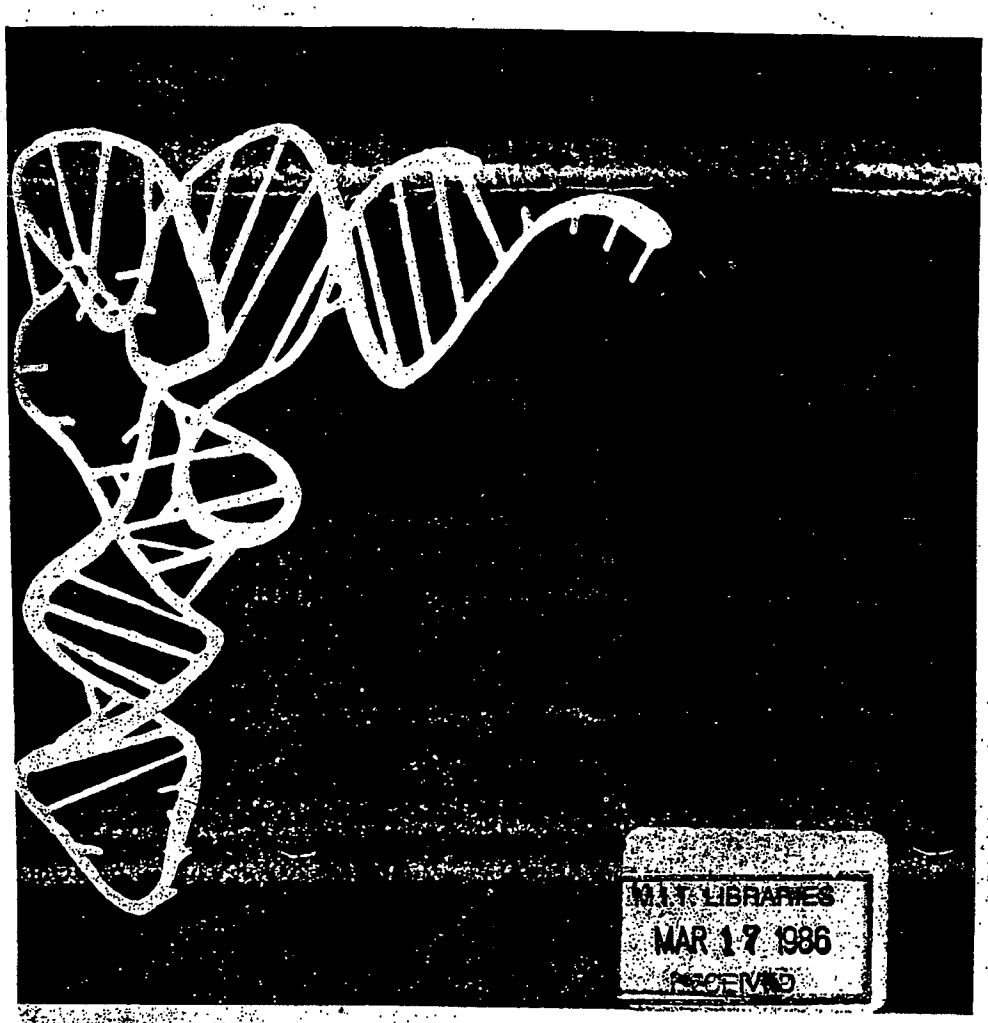
Exhibit C

Nucleic Acids Research



IRL PRESS
xford · Washington DC

ISSN 0305 1048
Coden NARHAD

Volume 14 Number 4 1986 Nucleic Acids Research

Evolution of the casein multigene family: conserved sequences in the 5' flanking and exon regions

Li-yuan Yu-Lee, Lizabeth Richter-Mann, Craig H.Couch, A.Francis Stewart[1+], Anthony G.Mackinlay[1] and Jeffrey M.Rosen*

Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA and [1]School of Biochemistry, University of New South Wales, Kensington 2033, Australia

Received 13 November 1985; Accepted 22 January 1986

ABSTRACT

The rat α- and bovine α_{s1}-casein genes have been isolated and their 5' sequences determined. The rat α-, β-, γ- and bovine α_{s1}-casein genes contain similar 5' exon arrangements in which the 5' noncoding, signal peptide and casein kinase phosphorylation sequences are each encoded by separate exons. These findings support the hypothesis that during evolution, the family of casein genes arose by a process involving exon recruitment followed by intragenic and intergenic duplication of a primordial gene. Several highly conserved regions in the first 200 base pairs of the 5' flanking DNA have been identified. Additional sequence homology extending up to 550 base pairs upstream of the CAP site has been found between the rat α- and bovine α_{s1}-casein sequences. Unexpectedly, the 5' flanking promoter regions are conserved to a greater extent than both the entire mature coding and intron regions of these genes. These conserved 5' flanking sequences may contain potential cis regulatory elements which are responsible for the coordinate expression of the functionally-related casein genes during mammary gland development.

INTRODUCTION

The caseins are a family of milk phosphoproteins which form calcium-dependent micelles (1). In the virgin gland, casein mRNAs are already detectable, although in differing amounts. During mammary gland development, these mRNAs become highly abundant, comprising 65% of total poly(A)$^+$RNA in the lactating gland. Such a dramatic increase in casein mRNA levels, ranging from several hundred- (α and β) to a thousand (γ) -fold, reflects both a low basal level in the uninduced virgin state as well as the proliferation of epithelial cells synthesizing caseins in the fully developed mammary gland (2). The expression of the α-, β- and γ-casein genes during mammary gland development is regulated in concert by steroid and peptide hormones (3), at both the transcriptional and post-transcriptional levels (4). These combined hormonal effects account for the large induction of casein mRNA levels in the induced gland.

As an initial attempt to elucidate the mechanism of coordinate regulation of the casein genes, the structure and organization of the casein mRNAs and the corresponding genomic sequences encoding these mRNAs have been studied.

Extensive nucleotide sequence analysis of a number of casein mRNAs from rat (5,6), mouse (7), guinea pig (8,9), and cow (10), has revealed a considerable divergence among the casein sequences, except for three regions which are highly conserved. These include the 5' non-coding, signal peptide and casein kinase phosphorylation sequences (6). Recent studies on the genomic structure and organization of the rat β- (11) and γ-casein (12) genes revealed that these three conserved structural gene regions are each encoded by separate exons. Thus, the evolution of the casein gene family involved initially the recruitment of exons with distinct functional domains to generate a primordial casein gene, which was followed by intragenic and intergenic duplications to generate the different members of the multigene family (11). Support for this hypothesis comes from genetic studies which show that the bovine casein genes occur as a gene cluster (13) and that all three mouse α-, β- and γ-casein genes are localized to a single chromosome (14).

In this paper, we report the isolation and characterization of the rat α- and bovine α_{s1}-casein gene 5' sequences. The 5' exon structure and organization of the rat α-, β- and γ-casein genes are found to be conserved, a finding supporting the previously proposed model of casein gene evolution. Furthermore, several regions in the 5' flanking DNA are highly conserved not only among the three rat genes but also between the rat α- and bovine α_{s1}-casein genes. The conserved sequences may represent cis-acting regulatory elements involved in the coordinate expression of the casein genes in response to developmental and hormonal signals.

MATERIALS AND METHODS

Phage Library Screening

Two rat genomic libraries, one a partial EcoRI (15) and the other a partial HaeIII, were obtained from Drs. T. Sargent, B. Wallace, and J. Bonner, and Drs. L. Jagodzinski and J. Bonner, respectively, and were screened as described previously (12) with an α-casein cDNA insert (6), to obtain the genomic rat α-casein clones. A bovine genomic library, constructed in λEMB3, was obtained from Dr. S. Ruppert (16) and was screened with a mixture of probes containing bovine α_{s1}-, α_{s2}-, β- and κ-casein and β-lactoglobulin cDNA inserts (10). The cDNA inserts were excised with PstI, then subjected to controlled digestion with Bal31 exonuclease to remove GC tails which otherwise caused unacceptably high backgrounds. Positives from the initial screen were plaque purified and identified by dot hybridization (17) using individual cDNA probes.

Plasmid Subcloning

DNA fragments from phage clones were subcloned into pBR325, pUC8 or pUC13 plasmids as described (12), or into M13mp8 or mp9 single-stranded phage as described by Messing (18) for further characterization such as restriction mapping, ligation, nick-translation, end-labeling and sequencing as previously described (12).

DNA Sequencing

End-labeled DNA was isolated either from low-melt agarose gel via Elutip-D columns as described (11) or by electroelution. Both Maxam-Gilbert (19) and dideoxy sequencing (18) methods were employed. All nucleic acid sequences reported were determined in both directions. Computer analysis of the sequence data was done with the HELEX Sequence Information System (C.B. Lawrence, personal communication). Dot matrix analysis was performed as described by Staden (20).

RESULTS

Rat α-Casein Gene

The size of the genomic rat α-casein gene was estimated by Southern blot analysis using an α-casein cDNA probe (6). A minimum size of 10-15 kb was obtained by BamHI and EcoRI digestions (data not shown). Several attempts to isolate the rat α-casein gene from a partial EcoRI rat library were unsuccessful, most likely due to the under-representation of the 15 kb EcoRI fragment in this amplified gene library. Using an independently-derived HaeIII library, an authentic rat α-casein gene clone, λα1, was isolated and further characterized (Fig. 2). The λα1 clone contains approximately 7.1 kb of 5' flanking DNA and 4.4 kb of the 5' portion of the α-casein gene. The first five exons comprising only 16% of the α-casein mRNA have been localized in a 5.1 kb EcoRI fragment by Southern blotting (data not shown) and direct DNA sequencing (Fig. 2B). The 5.1 kb EcoRI fragment most likely is part of the 15 kb EcoRI genomic fragment as its 3' end is composed of an artificial EcoRI site generated during library construction.

There appears to be only a single rat α-casein gene based on the analysis of genomic DNA blots (data not shown, 21). However, twelve other non-overlapping phage clones were isolated when the stringency of the hybridization conditions was lowered by reducing the temperature from 68° to 52° (data not shown). Since the hybridization signals from these clones are not observed at the higher temperature, these phage clones most likely contain non-casein genomic sequences which cross-hybridized with a repeated sequence

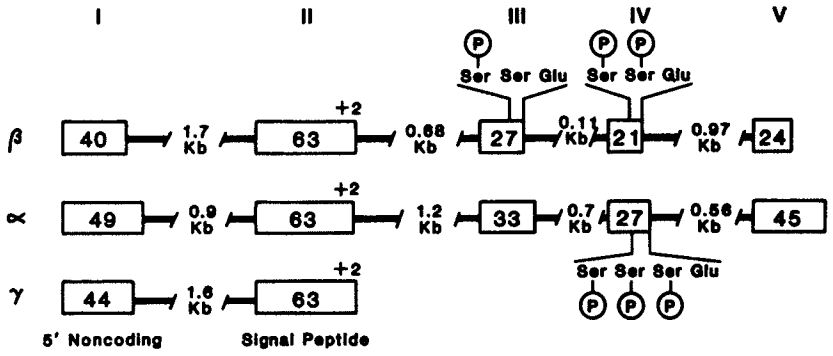


Figure 1. Conserved 5' exon structure of the rat casein genes. The structure of the exons (I-V) is compiled from the nucleotide sequence data of Jones et al. (11) for the β-casein gene, Fig. 2 of this paper for the α-casein gene, and Yu-Lee and Rosen (12) for the γ-casein gene. In addition, exon II of the γ-casein gene has been determined by dideoxy sequencing (data not shown). Numbers within boxes indicate exon sizes in bp and those in between the boxes represent approximate intron sizes in kb. The position of the second amino acid in the casein mRNA is indicated by +2. The position and the typical amino acid sequence of the phosphorylation sites (Ser-Ser-Glu) are as indicated where Ⓟ indicates that the residue is phosphorylated.

previously identified in the rat α-casein mRNA (6). This repeat is composed of 9 $^1/_2$ copies of an 18 bp repeat flanked by a duplicated sequence CCAA (6) and encodes 60 amino acids, thus accounting for the larger size of the rat α-casein (42 kDa) relative to the bovine α_{s1}- (23.6 kDa) (22) and other mammalian α-caseins, except perhaps the mouse α-casein (43 kDa) (7). Since the 5' portion of the rat α-casein gene represents possibly the more interesting region of the gene, it was compared with those of the rat β- (11), γ- and bovine α_{s1}-casein genes.

Conserved 5' Exon Structure

Previous analysis of the rat β-casein gene (11) revealed that the three conserved structural gene regions in the casein mRNA, i.e., the 5' noncoding, the signal peptide, and the phosphorylation site sequences, are encoded by separate exons. In the casein proteins, phosphorylation commonly occurs on a serine residue when it is in the sequence Ser-X-Glu or Ser-X-SerP, which constitutes a minor phosphorylation site (6). Addition of a second Glu codon to a (Ser)n-Glu sequence converts a minor to a major phosphorylation site, (Ser)n-Glu-Glu (6,11). In the rat β-casein gene, the second Glu codon is supplied by an RNA splicing event (11). Such an organization of exons is also found in the rat α- and γ-casein genes (Fig. 1). In all three rat genes, exon I encodes the 5' noncoding sequences, ending with AAG at the 3' terminus. A

similar 5' exon I structure is also found in the bovine α_{s1}- (Fig. 4B) and guinea pig α_{s2}-casein genes (L. Hall and R. Craig, personal communication). Exon II encodes the last 12 nucleotides of the 5' noncoding sequences, the entire 15 amino acid signal peptide sequence, and the first two amino acids of the mature casein proteins with a size of 63 bp in all three rat genes. Exon III in β contains a minor phosphorylation site, Ser-Ser-Glu, at the 3' terminus, while in α the sequence is Ser-Ser-Gln, which does not constitute a functional phosphorylation site but is clearly the remnant of a primordial site. Exon IV of both β and α contains several potentially phosphorylated serine residues ending in Ser-Ser-Glu while exon V of both genes begins with a Glu codon (see Jones et al. and Fig. 2B). The significance of this arrangement is evident when the two exons, IV and V, are joined by an RNA splicing event between the two glutamic acid codons. This process generates a major phosphorylation site sequence Ser-X-Ser-Ser-Glu-Glu in β and Ser-Ser-Ser-Glu-Glu in α. At the nucleotide level, the two Glu codons are invariably GAG-GAA where the underlined nucleotides form part of the consensus splice junction sequence (23) as previously described by Jones et al. (11). Thus, this unique exon arrangment has been preserved between the rat α and β genes. It seems then that these exons (III-V) as examined have evolved from an ancestral sequence containing a site for phosphorylation and calcium binding, in agreement with the proposed model of casein gene evolution (11). On the other hand, neither the size nor sequence of the introns has been conserved among the three rat genes.

Sequence Analysis of Rat and Bovine Casein Genes

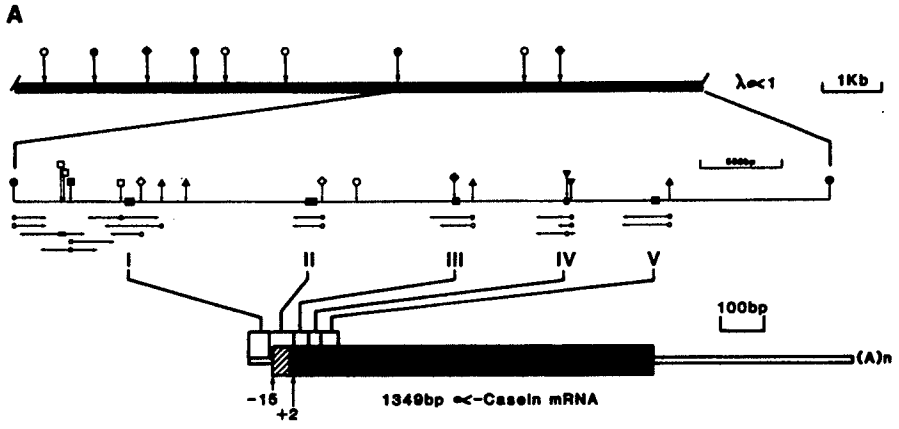
The 5' portions including 680 bp of flanking sequences of the rat α- (Fig. 2A and B), γ- (Fig. 3A and B) and bovine α_{s1}- (Fig. 4A and B) casein genes have been determined. All three genes share in the -30 bp region the unusual TATA sequence, TTTAAAT. This sequence which contains a T instead of an A in the second position, has been observed in the Goldberg-Hogness box of only a few other genes examined (24). An alternative though rare transcription initiation site is found in the bovine α_{s1}-casein gene (Fig. 4B). An aberrant minor α_{s1}-casein mRNA has been sequenced (10) whose 5' end begins at the -35 bp A residue and thus includes the conserved TATA sequence in its 5' noncoding region. A possible TATA sequence for the larger mRNA species can be found at about 50 bp upstream. Although the sequence TTTAAAT has been retained, the canonical 5'-G and 3'-AG nucleotides surrounding most eukaryotic TATA sequences (25) have been lost.

Several regions in the rat α-casein gene are found to contain alternating

purine-pyrimidine DNAs (Fig. 2B). The dinucleotide TG is repeated 38 times in the fourth intron, while the dinucleotide CA is repeated 39 times in the -650 bp region of the flanking DNA. The simple sequence $(CA)_3AA(CA)_2$ is found in the fifth intron, and is flanked by a duplicated sequence TCTTA, possibly the result of an earlier insertion event (26). These simple repeat sequences are not found in the rat β-, γ- or bovine α_{s1}- sequences determined to date.

Conserved 5' Flanking Sequences

To survey the relatedness of the 5' flanking sequences of the rat α-, β-, γ- and bovine α_{s1}-casein genes, pairs of 5' flanking sequences were compared by dot matrix analysis (Fig. 5) using the stringent criteria that 7 of 10 nucleotides must be identical to be scored positive. Several regions of conservation were found between the rat α and β sequences (Fig. 5A) as evidenced by the broken diagonal line extending to -200 bp. On the other hand, the rat α and γ sequences show considerable homology in the first exon and first 100 bp of flanking DNA (Fig. 5B). However, some of these homologies are displaced off the diagonal as a result of sequence rearrangements in the γ flanking DNA. No significant pattern of homology was observed beyond the first 200 bp among the three rat genes, although several regions of homology were evident between the α and β sequences at non-corresponding flanking positions (Fig. 5C) (see also Table 1B). The dark horizontal line is the result of homology between a $(CA)_4$ sequence at -670 bp in the β-casein gene (11) and a $(CA)_{39}$ sequence between -650 and -580 bp in the α gene (Fig. 2B). In contrast, a striking amount of homology was observed between the rat α- and bovine α_{s1}-casein sequences (Fig. 5D) extending up to -550 bp in the flanking DNA, through the conserved first exon and even extending several base pairs into the first intron. This comparison indicates that the rat α- and bovine



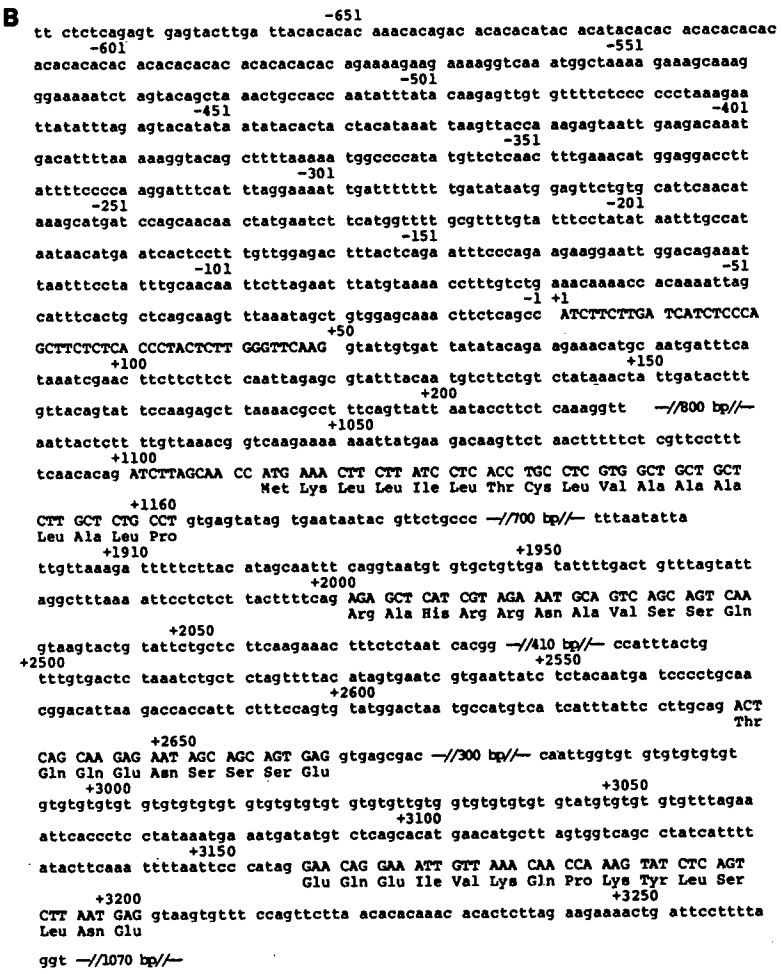
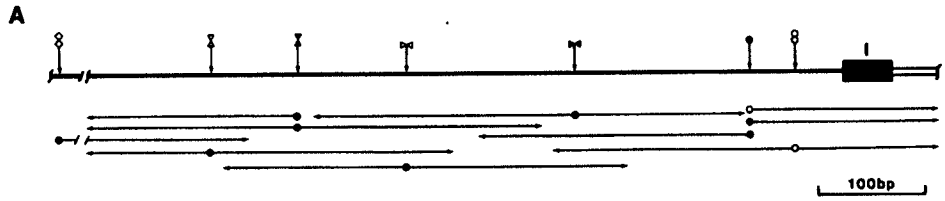


Figure 2. 5' sequences of the rat α-casein gene. A. Sequencing strategy. The solid line (top) represents the λα1 phage clone containing the 5' portion of the rat α-casein gene. A 5.1 kb EcoRI fragment (middle line) was subcloned, mapped to localize exon regions (boxes) (data not shown), and sequenced as indicated by arrows below the gene map. 5' end-labeling, solid circles; 3' end-labeling, open circles. The positions of the first five exons are indicated in the α-casein mRNA (bottom line). Noncoding sequences, open line; signal peptide sequence, hatched box; coding sequence, solid box. The restriction site symbols are as follows: (○) XbaI, (●) EcoRI, (◆) SstI, (□) DraI, (■) AvaII, (◇) TaqI, (△) HincII, (▲) HinfI, (▽) Fnu4HI, (▼) HpaII. B. The start of RNA transcription, +1, is determined by comparison with the published mRNA sequences established by primer extension and S1 mapping experiments (6). Exon sequences are shown in upper case letters and flanking and intron sequences in lower case letters. The encoded amino acids are shown below the nucleotide sequence. Gaps indicate sizes of unsequenced intron regions.



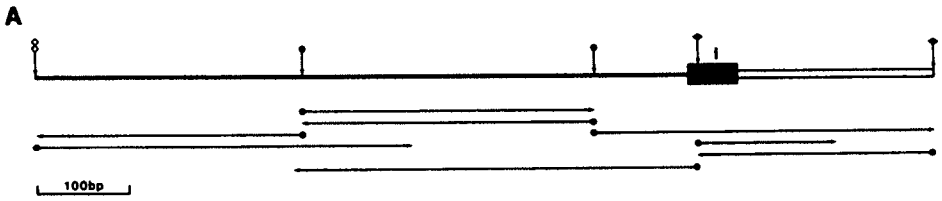
B

```
                                                -650
    tgccatgtt caattgacat ctctggaaga cttgctcttt tctgaaggga aatagagtag
                                -600
    agtggttcag gtgggagata gtgggatgga gggatgggaa gagtgaaagg agggaaaact
                          -550
    attgtcccag tgtattgtat gagagaagaa ttaaaaaaaa agagtccttc tgccgaggac
-500                                                        -450
    cctggtctgt tttgtaggac acagaggact tgtaactcta gctccattga atccaaatca
                                                -400
    cttttctgga ctttggactt tgaaggttcc tgcatgcaca tgatacacat acatgctctc
                                -350
    agggctacgc acatacacat gaaataaaat aatgaaataa agttcaaaaa tataaatgag
                          -300
    ataacaaata tctactgttc ataatctaat aacaaatggg gaaagttttg aaaacactag
                    -250
    aatttgagaa tgagttcata gaacaggtga gtctgccttc acacaaattc aattaaccac
-200                                                        -150
    cacacatctt ccattatctt atcatggcct caatcaaacg gtttaagaac tccctagaat
                                                -100
    ctgtggaaca aaatccagag agacaatttc taatgatatt gctttcttag aattcgaatg
                                -50
    tctttttagg tatttgaaac cacagaatta gcatatgatg ctagaacctg gtttaaatag
                -1 +1          +10          +20          +30          +40
    tgcgggagct acccactgct  ATCATCATCT ACCTATTCCT CTCGTCTTCC ACTTGGGAAG
                +50
    CAAG gtaatgtttc atattatttt tcttttatgg catacactaa aacatat
```

Figure 3. 5' sequences of the rat γ-casein gene. A. Sequencing strategy. The solid line represents 5' flanking DNA, solid box the first exon, and open line the first intron. Each arrow represents one sequencing gel. 5' end-labeling, solid circles; 3' end-labeling open circles. Symbols: (⊗) NcoI, (⧖) FokI, (⧗) Sau96A, (⋈) SphI, (⋈) XmnI, (⊗) NdeI.
B. The sequence of exon I and 200 bp of 5' flanking DNA have previously been published (12). The data is presented as in Fig. 2B.

α_{s1}-casein genes are more closely related to each other than the rat α is to the other rat casein genes. This finding extends the observation that the corresponding mRNA sequences are also more homologous to each other than each is to the other casein mRNAs in their respective species (10).

When the sequences of the first 200 bp of all three rat casein 5' flanking DNA were compared directly (Fig. 6), six regions showing good homology were observed. The first block (-23/-33) contains the TATA sequences which are flanked by the canonical 5'-G and 3'-AG nucleotides (25). The second (-51/-68) and third (-93/-103) blocks are AT rich and are the most highly conserved in all three rat genes, displaying an average of only 13% and 6% substitutions, respectively. The fourth (-113/-143), fifth (-150/-160) and sixth (-165/-174) blocks reveal greater similarity between the β and γ sequences as previously reported (11). For example, only 18% and 29% substitutions were observed in the fifth and fourth homology blocks, respectively. This may



B

```
                                                    -650
ctgcag tccatggggg tcacaaagaa ctggacacga cttagaaact aaacaacaac
                                                -600
aatttatatc agaatgaatg aactagttac cacaactagt acacccaaaa tgaacaaaaa
                                            -550
atagcttggt ggtataatta aaatgccacc aaaatttata caataattat attttctttt
                                        -500
tgcaggaaaa agattagacc acatataatg taacttattt cacaaggtaa ataattataa
                                                            -400
-450 taaataatat ggattaactg agttttaaaa ggtgaaataa ataatgaatt cttctcatgg
                                                -350
tcttgtatgt taataaaaat tgaaaaattt tgaagacccc attttgtccc aagaatttca
                                            -300
tttacaggta ttgaattttt caaaggttac aaaggaaatt ttattgatat aataaatgca
                                        -250
tgttctcata ataaccataa atctagggtt ttgttggggt ttttttgtt tgttaattta
                                    -200
gaacaatgcc attccatttc ctgtataatg agtcacttct ttgttgtaaa ctctccttag
                                                        -100
-150 aatttcttgg gagaggaact gaacagaaca ttgatttcct atgtgagaga attcttagaa
                                                -50
tttaaataaa cctggttggt taaactgaaa ccacaaaatt agcattttac taatcagtag
                                            -1+1
gtttaaatag cttggaagca aaagtctgcc ATCACCTTGA TCATCAACCC AGCTTGCTGC
                                        +50
TTCTTCCCAG TCTTGGGTTC AAG gtattatgta tacatataac aaaatttcta
                                    +100
tgattttcct ctgtctcatc tttcattctt cactaatacg cagttgtaac ttttctatgt
                                                        +200
    +150
gattgcaagt attggtactt tcctatgata tactgttagc ttaaaaatat atttgcaaat
                                                +250
gttgatacta tctatctcag agctataggt gaaaaattaa atacttttat aaagaccaaa
                                            +300
ttgatcattt tttaaacgaa aatttcttat atactga aaa tgtagataca taact
```

Figure 4. 5' sequences of the bovine α_{s1}-casein gene. A. Sequencing strategy is presented as in Fig. 3A. Symbols: (◇) PstI, (●) EcoRI, (◁▷) BclI.
B. The start of RNA transcription, +1, is located by homology with the rat α-casein sequence. It was previously identified as occurring two nucleotides further upstream on the basis of primer extension analysis (10). However, this result was probably inaccurate since the size standard was obtained from a sequencing ladder generated from a different region of the sequence. The data is presented as in Fig. 2B.

indicate that the two genes are more closely related to each other than to the rat α-casein gene. Interestingly, the sequence AGAATT is found in both the third and fifth blocks, where it also appears as an inverted repeat in the third block. Aside from these six blocks, few regions of extensive homology were observed. In contrast, the bovine α_{s1} sequence displays excellent homology with the rat α sequence, exhibiting an average of only 4.6% and 19%

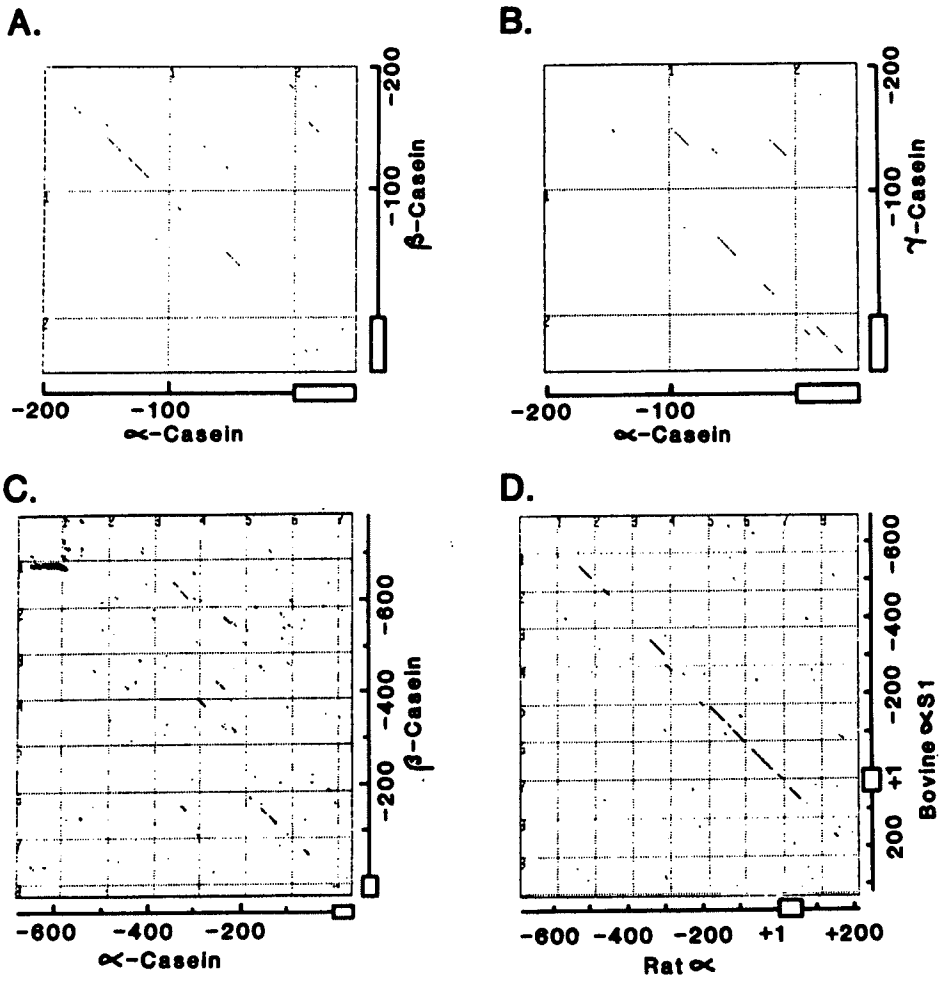
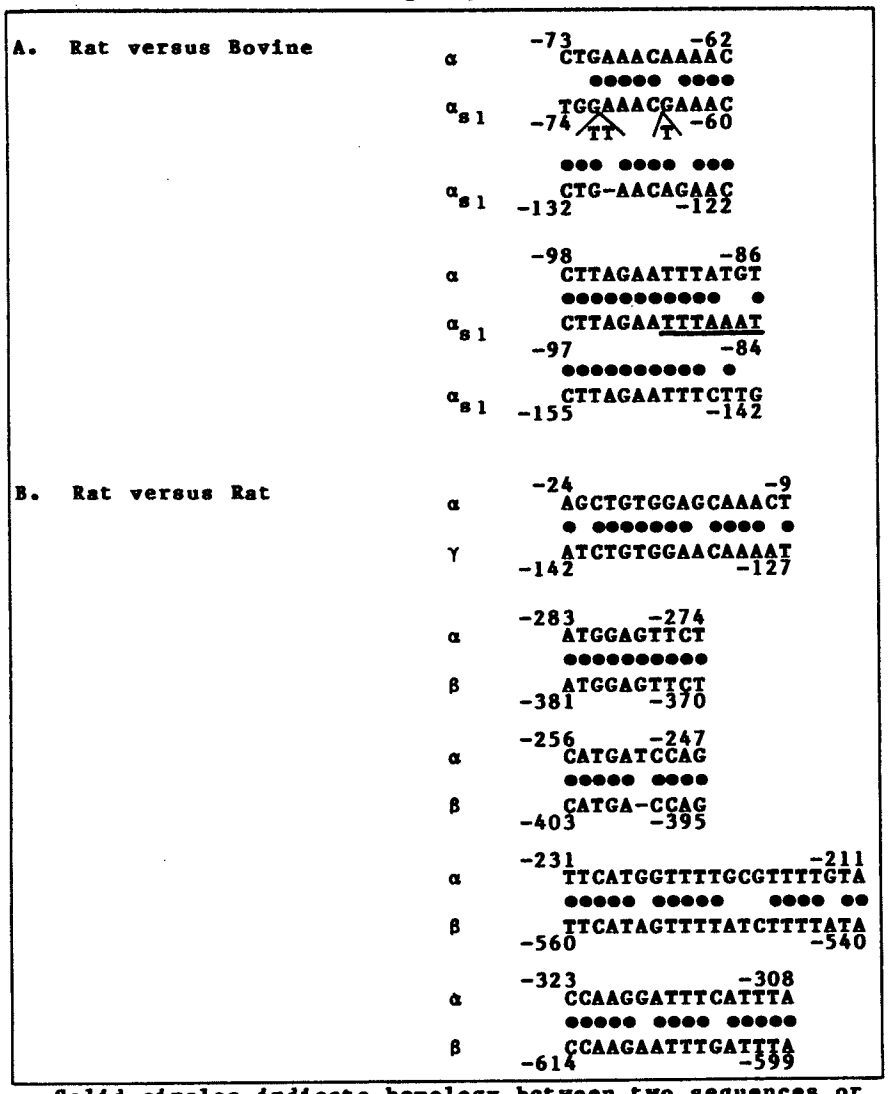


Figure 5. Dot matrix analyses of the rat α-, β-, γ- and bovine α_{s1}-casein 5' regions. The 5' regions of the casein genes being compared are indicated by the maps along the X and Y axes. Boxes represent exon I. Numbers refer to the relative distance in bp from +1. To be scored as positive, at least 7 out of 10 nucleotides are required to match. A. and C. Rat α vs rat β. B. Rat α vs rat γ. D. Rat α vs bovine α_{s1}.

substitutions in the first three and latter three homology blocks, respectively. For comparison, the first 200 sites of 5' flanking DNA showed an overall average of 25% substitution, while the entire mature protein coding region and the first 140 sites of intron I (excluding the splice junction consensus sequence) showed 36.5% (10) and 40% substitutions, respectively.

TABLE I.
Other 5' Flanking Sequence Homologies



Solid circles indicate homology between two sequences or between the first and third sequences. Numbers refer to their positions in the 5' flanking DNA. The solid line denotes the aberrant upstream promoter found in the bovine α_{s1}-casein gene (see Fig. 4B).

Thus, the conservation of flanking sequences is greater than those of both the coding and intron regions of the genes, suggesting that these regions have been selected for during evolution.

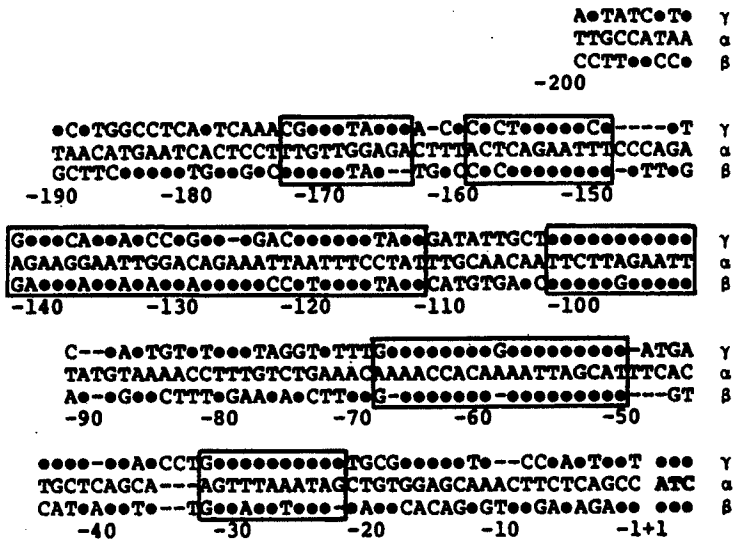


Figure 6. Conservation of the 5' flanking regions of the casein genes. The sequences of the rat α-, β- and γ-casein 5' flanking DNA were aligned to give maximum homology. +1 designates the start of exon I in α which is shown in bold-faced letters. All numbering refers to the α sequence. Dots indicate homology of the γ or β sequence relative to that of α. Gaps are introduced to maximize homology. The TATA sequence and five conserved regions are enclosed by boxes.
N.B. The start of transcription of the rat β-casein gene was previously estimated by primer extension sequence analysis (5). However, since a duplicated ATC sequence is present at the putative CAP site, the site of initiation could be at the 5' most ATC, making the size of exon I shown in Fig. 1 and Jones et al. (11) 43 instead of 40 bp.

DISCUSSION

Caseins represent one of the most rapidly diverging protein families studied (27) where the principal function of these milk proteins is to provide a super-saturating concentration of calcium, phosphates and essential amino acids to the young. Several classes of bovine caseins have been identified according to their physical and chemical properties: α_s-, β- and κ-caseins. While the α_s- and β-caseins are sensitive to calcium, the κ-caseins are not and are essential for the formation of micelles in the milk (1). The divergence of the bovine α_{s1}- and rat α-casein mRNAs involves extensive deletions/insertions in the protein coding regions (10), and in particular, an insertion into the rat α-casein mRNA of a 180 bp repeat element (6). The divergence between the bovine and rat β-casein mRNAs involves mainly a high rate of point mutations rather than insertions/deletions or major sequence rearrangements

(66), probably as a result of the functional constraints imposed on the β-caseins which have been shown to be important for curd formation (M.J. Pearse and A.G. Mackinlay, unpublished observation). On the other hand, the smaller rat γ-casein mRNA appears to have either undergone extensive deletions or have resulted from only a partial duplication of a common ancestral gene such that its sequence shares a similar organization with only the 3' coding region of the bovine α_{s2}-casein mRNA (66).

Analysis of the genes for the three rat caseins showed that they are quite different in size ranging from 7.5 kb for β-casein (11) to 15 kb for γ-casein (12) and about 10-15 kb for α-casein. However the exon structure at the 5' ends of these genes is conserved (Fig. 1). The 5' noncoding sequence and the signal peptide sequence, representing two of the most highly conserved regions in the casein mRNAs, are each encoded by separate exons. The conservation of the 5' noncoding exon may be related to the formation of potential secondary structures as suggested previously for the rat β- (5) and bovine α_{s2}- (66) casein mRNAs, which in turn may have a role in post-transcriptional regulation of casein synthesis. The exons encoding the signal peptide of all of the calcium-sensitive caseins examined show an invariant size of 63 bp, consistent with previous results which indicated that this is the most highly conserved region in the casein mRNAs and proteins. These signal peptides are all 15 amino acids in length and contain an invariant lysine residue in position 2 and a cysteine residue in position 8 (6). These residues have been suggested previously to play a role in the translocation (28), recognition and removal (29) of the signal peptides. Thus, the conservation of this exon may be related to the efficient secretion of the casein proteins into milk.

The split architecture of the major casein phosphorylation site, previously reported for the β-casein gene (11), has also been found in the α-casein gene (Figs. 1 and 2B). On closer examination, the generation of a major site by an RNA splicing event between two exons containing minor sites may be advantageous as a means of concentrating phosphate groups in localized regions in the casein proteins. Such clustering of phosphate residues may facilitate cooperative interactions between the casein phosphates and the calcium phosphates of the casein micelles (30). Exons III and IV in the α-, as well as exons III-VI in the β-casein genes (11), all represent duplicated units of an ancestral exon containing the site of phosphorylation and calcium binding. On the other hand, no information is available on the exon structure of the hydrophobic domain of the α- or γ-casein genes. This domain in the β-

casein gene is encoded by one large exon (11) whose sequence has been the least conserved (5). Thus, the modern-day caseins appear to have been generated by exon shuffling (31) which brought together the various present-day functional domains. Other striking examples of gene evolution via exon shuffling include the light density lipoprotein receptor (32) and the calcium-dependent protease genes (33).

Two stretches of alternating purines and pyrimidines, $(CA)_{39}$ and $(TG)_{38}$, are found in the 5' -650 bp flanking region and fourth intron, respectively, of the α-casein gene (Fig. 2B). Since there are 10^5 copies of the $(TG)_{20-60}$ elements interspersed throughout the rat genome, it is perhaps not surprising to find these repeats in and around the α gene. A number of mammalian genes also contain (TG)n either in the flanking (34-37) or intron (38) regions. Since poly(dT-dG) (poly dC-dA) sequences have the potential to form Z DNA (39,40) and may be involved in enhancing gene transcription (41), the presence of the $(CA)_{39}$ and $(TG)_{38}$ repeats in the rat α-casein gene may be related to the higher basal level of α-casein mRNA relative to β and γ as observed in the virgin mammary gland (2) and in mammary explant cultures in the absence of lactogenic hormones (2).

The simple repeat $(AG)_{33}$ is found in the 5' - 750 bp flanking region of the β-casein gene (11). Such a sequence located 2 kb downstream from the rat somatostatin gene (34) resulted in extreme S1 nuclease sensitivity. The 5' location of the $(AG)_{33}$ sequence may thus confer S1 nuclease sensitivity to the 5' end of the rat β-casein gene, but this remains to be established experimentally. S1 nuclease sensitivity has been observed near the transcriptional start sites of actively transcribed genes (42,43). These simple repeats are not preserved among the rat casein genes, thus they may underlie the differences in their basal levels of gene expression.

It is interesting to note that the TATA sequence in the rat β-casein gene is TATATA (11) while that observed in both α- and γ- as well as bovine α_{s1}-casein is TTTAAAT (Figs. 2B, 3B and 4B). The transition of an A to a C, T, or G in the second position of the TATA sequence has been demonstrated to reduce the *in vitro* efficiency of the promoter sequence in other genes (44). In addition, preliminary *in vitro* transcription (L.-Y. Yu-Lee and S.Y. Tsai, unpublished observation) and transient expression (C.A. Bisbee, unpublished observation) analyses have demonstrated that the TTTAAAT sequence is a weak promoter. These observations are unexpected considering the relative abundance of these mRNAs in the lactating gland, and suggest that the intrinsic long half-lives rather than exceptional transcription rates may account for the marked accumulation of these mRNAs. The upstream promoter (-84/-90) for

the infrequent bovine α_{s1}-casein transcript (Fig. 4B) is located in a stretch of DNA that appears to have been duplicated. The -61 to -120 bp region shares homology with the -121 to -180 bp flanking DNA and this duplication although still apparent, is less well-preserved in the rat flanking DNA. For example, the rat α-casein sequence CTTAGAATTT (-98/-89) is homologous to two bovine sequences at -97 bp (corresponding position) and -155 bp (duplicated upstream position) (Table 1A). Interestingly, the α_{s1}-casein upstream promoter may have been generated after the initial duplication event since it is not found in either the corresponding rat nor duplicated bovine flanking DNA. The rat α-casein sequence at -73 bp which shares homology with the two bovine α_{s1}-casein sequences at -74 and -132 bp, defines one boundary of the duplicated region (Table 1A). Furthermore, the α_{s1}-casein sequence at -74 bp, GTTGGTTAAACT, resembles the CAAT sequence found in rabbit β-globin gene, GTTGGCCAATCT (-80/-69) (45). The other copy at -132 bp has diverged such that it no longer resembles the canonical CAAT sequence (25). Nevertheless, it is interesting to speculate that the upstream TTTAAAT sequence together with the sequence at -132 bp may have allowed the transcription of the infrequent upstream α_{s1}-casein gene transcript.

The rat and cow diverged about 75 million years ago, around the time of the mammalian radiation (27). Based on the average rate of 5.37×10^{-9} mutations/site/year for nucleotide changes at silent codon or intron positions (46), the number of changes predicted for the first 5' flanking 200 bp is 81 bp while only 53 are observed. On closer examination, the 5' flanking region distal to the rat and bovine α-casein genes showed 28% substitutions (28/100 sites compared) and the region proximal to the genes showed 21% substitutions (21/100 sites compared). Both rates of substitutions are unusual since they are lower than those reported for six different sets of rat/mouse versus human genes, where the distal and proximal 5' flanking regions as well as the introns are evolving neutrally with similar rates (average of 37% substitutions) (47). The constraint observed in the evolution of the casein promoter-associated flanking DNAs is more striking when only those regions within the six homology blocks (see below) are considered. Therefore, the primordial casein gene duplication unit may have included at least the first 200 bp of 5' flanking DNA and this flanking region has been conserved to a greater extent than even the mature protein coding regions of the casein genes. As further support, the first 90 bp of 5' flanking sequences of the guinea pig α_{s2}-casein gene have been found to display excellent homology with the corresponding rat γ-casein sequence (L. Hall and R. Craig, personal communication).

Although a direct linkage has yet to be established for the casein genes, they occur as a gene cluster (13) and have been mapped to a single chromosome (14). The three rat casein genes have been shown to be regulated in a coordinate manner during mammary gland development and hormonal induction in explant cultures (2). Each casein gene may, therefore, possess common cis-acting elements which direct their synchronous response to these hormonal signals and other trans-acting regulatory molecules. Analysis of the rat and bovine genes (Figs. 2, 3, and 4) revealed that in addition to the TATA sequences, there are five well-conserved regions of sequence homology in the first 200 bp of the 5' flanking regions (Fig. 6). These conserved blocks reside at similar flanking positions, average 11 to 31 nucleotides long, and are AT rich in contrast to the GC rich nature of other promoter region DNAs (48,49). Two of the blocks (third and fifth) are similar in sequence, (A/T)CT (C/T) AGAATT. Interestingly, a third homologous sequence CCCAAGAATT is found in addition at -330 bp in the bovine α_{s1} gene (Fig. 4B). The sequence TGTT, found in the sixth homology block, has also been observed in the -200 bp region of the flanking DNA of several androgen-regulated rat genes (50). Other blocks of sequence homologies between gene pairs averaging 10-21 bp long (Table 1B) are observed further upstream but these do not reside at corresponding flanking positions. These conserved sequences do not resemble hormone receptor binding sites (see below) and their functions are not known. However, they may be analogous to the short (9-24 bp) repetitive sequences which have been demonstrated to be required for the coordinate induction of several sets of genes (51), for example, the yeast amino acid synthesis genes (52) and Drosophila heat shock genes (53). Alternatively, these sequences may be involved in the tissue-specific expression of the casein genes in the mammary gland.

Glucocorticoids are involved in the induction of casein gene expression (2) while progesterone inhibits casein synthesis during pregnancy (54). Estrogen, on the other hand, is involved in the proliferation of the mammary epithelial cells during mammary gland development (3). Accordingly, the 5' flanking sequences were examined for sequence similarities with reported rat glucocorticoid (55), rabbit progesterone (56) and chicken estrogen (57) receptor binding sites. The hexanucleotide TGT(T/C)CT shown to be part of a number of glucocorticoid receptor binding sites by DNase footprinting (55,56,58,59, 60) and transfection analysis (59,60), is seen several times in the casein flanking DNA. For example, it is found at positions -360 in the rat α gene (Fig. 2B), -230, -480 in the rat γ gene (Fig. 3B), and -120, -210, -270 in the

bovine α_{s1} gene (Fig. 4B). In the α gene, the hexanucleotide is part of a sequence TGGCCCCATATGTTCT which shares good homology (underlined) with the consensus sequence for binding of the glucocorticoid receptor, (T/C)GGTN(A/T) CA(A/C)(A/T)NTGT(T/C)CT derived from the binding sites in the MMTV LTR (55), human metallothionein (59), rabbit uteroglobin (58), and chicken lysozyme (56, 60) genes. Interestingly, in the γ gene, the complement of the -230 hexanucleotide is part of a sequence GAGTTCATAGAACA which shares homology (underlined) with the estrogen receptor binding sequence identified in the chicken vitellogenin II gene (57), GAGCTGAAAGAACAC. The other γ gene hexanucleotide at -480 follows immediately a sequence AGTCCTCTGTGTCCT which shares good homology (underlined) with two sequences identified in the chicken lysozyme gene (56,60), ATTCCTCTGT, which bind the heterologous rabbit progesterone receptor. This hexanucleotide is also found in the 5' noncoding sequence of both the rat and bovine β-casein mRNAs (25, 66), and the rat sequence has the potential to form a stem-loop secondary structure. Finally, several sequences sharing homology (underlined) with the progesterone receptor binding sequence (56,60) have been identified near the bovine α_{s1} gene (Fig. 4B). These are TTTCCTTTGT at -300, ATTTCCTATGT at -110, and TTTCCTCTGT at +90, which is an intron sequence. Internal hormone receptor binding sites have been reported for the human growth hormone (61) and MMTV genes (62). It is interesting that some of these potential hormone responsive sequences are found to be clustered in the same region, even share similar sequences, or they are inverted relative to the genes. The physiological significance of these observations for the human growth hormone (61) and MMTV genes (62). It is interesting that some of these potential hormone responsive sequences are found to be clustered in the same region, even share similar sequences, or they are inverted relative to the genes. The physiological significance of these observations is not clear. In any event, the functional significance of these sequence homologies remains to be established by directed mutagenesis and transfection analysis.

In conclusion, these studies extend the characterization of the structure and organization of the rat casein gene family. The conservation of the 5' exon structures provide strong support for the hypothesis that the casein gene family arose by recruitment of exons containing functional domains followed by intragenic duplication at about the time of the appearance of the primitive mammals, approximately 300 million years ago (11). Intergenic duplication then generated the members of the casein gene family, before the mammalian radiation approximately 75 million years ago. The striking conservation of blocks

of 5' flanking sequences points to their potential functional role, for example, in the coordinate induction or depression of these genes during mammary gland development and/or in tissue-specific expression. This does not preclude the possibility that intragenic and/or 3' flanking sequences are also important for casein gene regulation as have recently been reported for other genes (63, 64, 65). The functional role of these sequences is currently being investigated by transfection experiments using an entire rat β-casein gene, a truncated rat α-casein minigene, and several fusion gene constructions.

ACKNOWLEDGMENTS

We thank Sara Rupp for technical assistance, and Patricia Kettlewell for typing the manuscript. This work was supported by American Cancer Society grant BC425 to L.-Y. Y.-L., Public Health Service grant CA16303 from the National Institutes of Health to J.M.R., and grants from the Australian Research Grants Scheme and Australian Dairy Research Committee to A.G.M. A.F.S. was the recipient of a Commonwealth Postgraduate Award.

*To whom correspondence should be addressed
+Present address: Institute of Cell and Tumor Biology, German Cancer Research Center, D-6900 Heidelberg, FRG

REFERENCES

1. Waugh, D.F. (1971) in Milk Proteins, McKenzie, H.A., Ed., Vol. II, p. 3-85, Academic Press, New York.
2. Hobbs, A.A., Richards, D.A., Kessler, D.J. and Rosen, J.M. (1982) J. Biol. Chem. 257, 3598-3605.
3. Topper, Y.J. (1970) Recent Prog. Horm. Res. 26, 286-308.
4. Guyette, W.A., Matusik, R.J. and Rosen, J.M. (1979) Cell 17, 1013-1023.
5. Blackburn, D.E., Hobbs, A.A. and Rosen, J.M. (1982) Nucl. Acids Res. 10, 2295-2307.
6. Hobbs, A.A. and Rosen, J.M. (1982) Nucl. Acids Res. 24, 8079-8098.
7. Hennighausen, L.G. and Sippel, A.E. (1982) Eur. J. Biochem. 125, 131-141.
8. Hall, L., Laird, J.E. and Craig, R.K. (1984) Biochem. J. 222, 561-570.
9. Hall, L., Laird, J.E., Pascall, J.C. and Craig, R.K. (1984) Eur. J. Biochem. 138, 585-589.
10. Stewart, A.F., Willis, I.M. and Mackinlay, A.G. (1984) Nucl. Acids Res. 12, 3895-3907.
11. Jones, W.K., Yu-Lee, L.-Y., Clift, S.M., Brown, T.L. and Rosen, J.M. (1985) J. Biol. Chem. 260, 7042-7050.
12. Yu-Lee, L.-Y. and Rosen, J.M. (1983) J. Biol. Chem. 258, 10794-10804.
13. Matyukov, V.S. and Urnyshev, A.P. (1980) Genetika 16, 884-886.
14. Gupta, P., Rosen, J.M., D'Eustachio, P. and Ruddle, F.H. (1982) J. Cell Biol. 93, 199-204.
15. Sargent, T.D., Wu, J.-R., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A. and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 3256-3260.
16. Ruppert, S., Scherer, G. and Schutz, G. (1984) Nature (London) 308, 554-557.

17. Kafatos, F.C., Jones, C.W. and Efstratiadis, A. (1979) Nucl. Acids Res. 7, 1541-1552.
18. Messing, J. (1983) in Methods in Enzymology, Grossman, L. and Moldave, K. Eds., Vol. 101, p. 20-78, Academic Press, New York.
19. Maxam, A.L. and Gilbert, W. (1980) in Methods in Enzymology, Grossman, L. and Moldave, K. Eds., Vol. 65, p. 499-560, Academic Press, New York.
20. Staden, R. (1982) Nucl. Acids Res. 10, 2951-2961.
21. Johnson, M.L., Levy, J., Supowit, S.C., Yu-Lee, L.-Y. and Rosen, J.M. (1983) J. Biol. Chem. 258, 10805-10811.
22. Mercier, J.-C, Grosclaude, F. and Ribadeau-Dumas, B. (1971) Eur. J. Biochem. 23, 41-51.
23. Mount, S.M. (1982) Nucl. Acids Res. 10, 459-472.
24. Campbell, S.M., Rosen, J.M., Hennighausen, L.B., Strech-Jurk, U. and Sippel, A.E. (1984) Nucl. Acids Res. 12, 8685-8697.
25. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
26. Calos, M.P. and Miller, J.H. (1980) Cell 20, 579-595.
27. Dayhoff, M.O. (1976) in Atlas of Protein Sequence and Structure, Dayhoff, M.O., Ed., Vol. 5, Suppl. 2, National Biomedical Research Foundation, Bethesda.
28. Inouye, S., Soberon, X., Franceschini, T., Nakamura, K., Itakura, K. and Inouye, M. (1982) Proc. Natl. Acad. Sci. USA 79, 3438-3441.
29. Walter, P. and Blobel, G. (1980) Proc. Natl. Acad. Sci. USA 77, 7112-7116.
30. Sleigh, R.W., Sculley, T.B. and Mackinlay, A.G. (1979) J. Dairy Res. 46, 337-342.
31. Gilbert, W. (1978) Nature (London) 271, 501.
32. Sudhof, T., Goldstein, J.L., Brown, M.S. and Russell, D.W. (1985) Science 228, 815-822.
33. Ohno, S., Yasufumi, E., Imajoh, S., Kawasaki, H., Kisaragi, M. and Suzuki, K. (1984) Nature 312, 566-570.
34. Hayes, T.E. and Dixon, J.E. (1985) J. Biol. Chem. 260, 8145-8156.
35. Maurer, R.A. (1985) DNA 4, 1-9.
36. Searle, P.F., Davison, B.L., Stuart, G.W., Wilkie, T.M., Norstedt, G. and Palmiter, R.D. (1984) Mol. Cell. Biol. 4, 1221-1230.
37. Suwa, Y., Mizukami, Y., Sogawa, K. and Fujii-Kuriyama, Y. (1985) J. Biol. Chem. 260, 7980-7984.
38. Qasba, P.K. and Safaya, S.K. (1984) Nature (London) 308, 377-380.
39. Hamada, H., Pretrino, M.G., Kakunaga, T., Seidman, M. and Stollar, B.D. (1984) Mol. Cell. Biol. 4, 2610-2621.
40. Nordheim, A. and Rich, A. (1983) Nature (London) 303, 674-679.
41. Hamada, H., Seidman, M., Howard, B.H. and Gorman, C.M. (1984) Mol. Cell. Biol. 4, 2622-2630.
42. Larson, A. and Weintraub, H. (1982) Cell 29, 6909-622.
43. McKnight, S.L. (1982) Cell 31, 355-356.
44. Concino, M., Goldman, R.A., Caruthers, M.H. and Weinmann, R. (1983) J. Biol. Chem. 28, 8493-8496.
45. Dierks, P., van Ooyen, A., Mantei, N. and Weissman, C. (1981) Proc. Natl. Acad. Sci. USA 78, 1411-1415.
46. Miyata, T., Hayashida, H., Kikuno, R., Hasegawa, M., Kobayashi, M. and Koike, K. (1982) J. Mol. Evol. 19, 28-35.
47. Soares, M.B., Schon, E., Henderson, A., Karathanasis, S.K., Cate, R., Zeitlin, S., Chirgwin, J. and Efstratiadis, A. (1985) Mol. Cell. Biol. 5, 2090-2103.
48. Dynan, W.S. and Tjian, R. (1983) Cell 35, 79-87.
49. McKnight, S.L., Kingsbury, R.C., Spence, A. and Smith, M. (1984) Cell 37, 253-262.

50. Williams, L., McDonald, C. and Higgins, S. (1985) Nucl. Acids Res. 13, 659-672.
51. Davidson, E.R., Jacobs, H.T. and Britten, R.J. (1983) Nature (London) 301, 468-470.
52. Donahue, T.F., Daves, R.S., Lucchini, G. and Fink, G.R. (1983) Cell 32, 89-98.
53. Pelham, H.R.B. (1982) Cell 30, 517-528.
54. Rosen, J.M., O'Neal, D.L., McHugh, J.E. and Comstock, J.P. (1978) Biochemistry 17, 290-297.
55. Scheidereit, C., Geisse, S., Westphal, H.M. and Beato, M. (1983) Nature (London) 304, 749-752.
56. von der Ahe, D., Janich, S., Scheidereit, C., Renkawitz, R., Schutz, G. and Beato, M. (1985) Nature (London) 313, 706-709.
57. Jost, J.-P., Seldran, M. and Geiser, M. (1984) Proc. Natl. Acad. Sci. USA 81, 429-433.
58. Cato, A.C.B., Geisse, S., Wenz, M., Westhal, H.M. and Beato, M. (1984) EMBO J. 3, 2771-2778.
59. Karin, M., Haslinger, A., Holtgreve, H., Richards, R.I., Krauter, P., Westphal, H.M. and Beato, M. (1984) Nature (London) 308, 513-519.
60. Renkawitz, R., Schutz, G., von der Ahe, D. and Beato, M. (1984) Cell 37, 503-510.
61. Slater, E.P., Rabenau, O., Karin, M., Baxter, J.D., and Beato, M. (1985) Mol. Cell. Biol. 5, 2984-2992
62. Payvar, F., DeFranco, D., Firestone, G.L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J.-A. and Yamamoto, K.R. (1983) Cell 35, 381-392.
63. Charnay, P., Treisman, R., Mellon, P., Chao, M., Axel, R., and Maniatis, T. (1984) Cell 38, 251-263.
64. Merrill, G. F., Hauschka, S.D., and McKnight, S.L. (1984) Mol. Cell. Biol. 4, 1777-1784.
65. Grosschedl, R. and Baltimore, D. (1985) Cell 41, 885-897.
66. Stewart, A.F., Beattie, C.W., Bonsing, J., Shah, F., Willis, J.M., and Mackinlay, A.G. (1986) Mol. Biol. Evol, in press.